Kernel eigen analysis

Manel Martínez-Ramón

ECE, UNM

October 9, 2017

Principal component analysis in $\mathbb{R}^D$

- Assume that a set of data $\{\mathbf{x}_1 \cdots \mathbf{x}_N\} \in \mathbb{R}^D$ with zero mean is available.
- Its autocorrelation function can be estimated as

$$\mathbf{R} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top = \mathbf{X}\mathbf{X}^\top$$

- This matrix has a representation in terms of eigenvectors and eigenvalues of the form

$$\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$$

where $\mathbf{\Lambda}$ is a diagonal matrix whose elements $\mathbf{\Lambda}_{ii} = \lambda_i$ are the matrix eigenvalues, and $\mathbf{Q}$ is a matrix whose column contain orthonormal vectors called eigenvectors.

Principal component analysis in $\mathbb{R}^D$

- The construction of eigenvectors is based on the criterion of minimum mean square error projection.
- The idea is to find a direction in the space where the projected data has the minimum mean square error with respect to the original data. The projection for an element is

$$\hat{\mathbf{x}}_n = \langle \mathbf{x}_n, \mathbf{q} \rangle \, \mathbf{q}$$

- The projection error is

$$\mathbf{x}_n - \hat{\mathbf{x}}_n = \mathbf{x}_n - \langle \mathbf{x}_n, \mathbf{q} \rangle \, \mathbf{q}$$

The PCA theorem

- Theorem: A set of N vectors of dimension D are to be modelled using L orthogonal basis vectors $\mathbf{q}_n$ and scores $\mathbf{z}_n$. The reconstruction error is

$$J(\mathbf{Q}, \mathbf{Z}) = \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{x}n - \mathbf{Q}\mathbf{z}_n||^2$$

The minimal reconstruction error is achieved if the basis $\mathbf{Q}$ contains the L largest eigenvectors of the empirical covariance matrix of the data.

- Proof:
  The one dimensional solution has the reconstruction error

$$
\begin{aligned}
J(\mathbf{q}_1, \mathbf{z}_1) =& \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{q}_1 z_{n,1}||^2 \\
=& \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{x}_n||^2 - 2 z_{n,1} \mathbf{q}_1^\top \mathbf{x}_n + z_{n,1}^2 \mathbf{q}_1^\top \mathbf{q}_1 \\
& \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{x}_n||^2 - 2 z_{n,1} \mathbf{q}_1^\top \mathbf{x}_n + z_{n,1}^2
\end{aligned}
$$

  To minimize it, we need to compute its derivative wrt $z_{n,1}$

- Proof (cont.):

$$\frac{d}{dz_{n,1}} J(\mathbf{q}_1, \mathbf{z}_1) = \frac{d}{dz_{n,1}} \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{x}_n||^2 - 2z_{n,1}\mathbf{w}_1^\top \mathbf{x}_n + z_{n,1}^2$$

$$= \frac{1}{N} \left(-2\mathbf{q}_1^\top \mathbf{x}_n + 2z_{n,1} \right)$$

Nulling the derivative leads to

$$z_{n,1} = \mathbf{q}_1^\top \mathbf{x}_n$$

whose reconstruction error is

$$J(\mathbf{w}_1, \mathbf{z}_1) = \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{x}_n||^2 - 2\mathbf{x}_n^\top \mathbf{q}_1 \mathbf{q}_1^\top \mathbf{x}_n + z_{n,1}^2$$

$$= \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{x}_n||^2 - z_{n,1}^2$$

- Proof (cont.):
  Now, this error has to be minimized with the constraint $\mathbf{q}_n^\top \mathbf{q}_n = 1$.
  Then, applying Lagrange optimization we have to minimize the functional

$$\begin{aligned} L(\mathbf{q}_i) &= -\frac{1}{N}\sum_{n=1}^{N} \mathbf{q}_1^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{q}_1 + \lambda_1(\mathbf{q}_1^\top \mathbf{q}_1 - 1) \\ &= -\mathbf{q}_1^\top \hat{\mathbf{R}} \mathbf{q}_1 + \lambda_1(\mathbf{q}_1^\top \mathbf{q}_1 - 1) \end{aligned}$$

  Taking derivatives gives $\hat{\mathbf{R}}\mathbf{q}_1 = \lambda_1 \mathbf{q}_1$
  Hence, λ_1 and $\mathbf{q}_1$ are, respectively, an eigenvalue and an eigenvector of the autocorrelation matrix.

- Now we use the nonlinear transformation $\boldsymbol{\varphi}(\mathbf{x})$ into an RKHS with kernel $k(\cdot,\cdot)$.
- Given the previous set $\mathbf{x}_n$, we can construct a matrix $\boldsymbol{\Phi}$ of transformed data.
- Its autocorrelation matrix is

$$\mathbf{C} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\top = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$$

and we know that

$$\mathbf{C}\mathbf{V} = \boldsymbol{\Lambda}\mathbf{V}$$

- Now we assume that the eigenvectors are a linear combination of the data

$$\mathbf{V} = \boldsymbol{\Phi}\boldsymbol{A}$$

- Expressing C in terms of the data matrix gives

$$\mathbf{C}\mathbf{V} = \frac{1}{N}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\mathbf{V} = \boldsymbol{\Lambda}\mathbf{V}$$

and with $\mathbf{V} = \boldsymbol{\Phi}\boldsymbol{A}$

$$\frac{1}{N}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{A} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{A}$$

Now we premultiply by $\boldsymbol{\Phi}^\top$

$$\frac{1}{N}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{A} = \boldsymbol{\Lambda}'\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{A}$$

which equals to

$$\frac{1}{N}\mathbf{K}^2\boldsymbol{A} = \boldsymbol{\Lambda}'\mathbf{K}\boldsymbol{A}$$

- Finally

$$\mathbf{K}\boldsymbol{A} = N\boldsymbol{\Lambda}'\boldsymbol{A}$$

- This results say that matrix $\boldsymbol{A}$ containing vectors $\boldsymbol{\alpha}_k$ is the set of eigenvectors of $\mathbf{K}$, and its eigenvalues are $N\boldsymbol{\Lambda}'$, which is a matrix containing the nonzero eigenvalues of C scaled by N.
- The final result can be summarized as follows:
  1. If $\boldsymbol{\alpha}_k$ is an eigenvector of the kernel matrix $\mathbf{K}$, then $\mathbf{v}_k = \boldsymbol{\Phi}\boldsymbol{\alpha}_k$ is an eigenvector of $\mathbf{C}$.
  2. If λ_k is the eigenvalue of $\mathbf{v}_k$, then $N\lambda_k$ is the eigenvalue of $\boldsymbol{\alpha}_k$.

Centering the data around the origin

- The previous result assumes that the data is centered around the origin. This is in general not guaranteed in the feature space regardless of the distribution of the data in the input space.
- In order to center the data we need to compute the mean and subtract it from all vectors:

$$\bar{\boldsymbol{\varphi}}(\mathbf{x}_n) = \boldsymbol{\varphi}(\mathbf{x}_n) - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\varphi}(\mathbf{x}_i) = \boldsymbol{\varphi}(\mathbf{x}_n) - \frac{1}{N}\boldsymbol{\Phi}\mathbf{1}_N$$

- The centred matrix can be written as

$$\tilde{\boldsymbol{\Phi}} = \boldsymbol{\Phi} - \frac{1}{N}\boldsymbol{\Phi}\mathbf{1}_N\mathbf{1}_N^\top = \boldsymbol{\Phi} - \frac{1}{N}\boldsymbol{\Phi}\mathbf{1}_{N,N}$$

where $\mathbf{1}_N^\top$ is a row of N 1's, and $\mathbf{1}_{N,N}$. is a matrix of ones of dimension N.

- The new kernel matrix is, straigthforwardly

$$\tilde{\mathbf{K}} = K + \frac{1}{N^2}\mathbf{1}_{N,N}\mathbf{K}\mathbf{1}_{N,N} - \frac{1}{N}\mathbf{1}_{N,N}\mathbf{K} - \frac{1}{N^2}\mathbf{K}\mathbf{1}_{N,N}$$

- Homework: Substract the mean of a (training) data set to a another (test) data set.

- The approximation made by projecting a vector $\boldsymbol{\Phi}(\mathbf{x})$ over an eigenvector $\mathbf{v}_k$ is

$$\tilde{\boldsymbol{\Phi}}(\mathbf{x}) = \langle \boldsymbol{\Phi}(\mathbf{x}), \mathbf{v_k} \rangle \mathbf{v_k}$$

  Since $\mathbf{v}_k = \boldsymbol{\Phi}\boldsymbol{\alpha}_k$ then

$$\langle \boldsymbol{\Phi}(\mathbf{x}), \mathbf{v_k} \rangle = \boldsymbol{\Phi}^\top(\mathbf{x})\boldsymbol{\Phi}\boldsymbol{\alpha_k} = \mathbf{k}^\top(\mathbf{x})\boldsymbol{\alpha_k}$$

  where $\mathbf{k}(\mathbf{x})$ is the vector of kernel products $k(\mathbf{x}, \mathbf{x}_n)$

- The projection error is then

$$\|\boldsymbol{\Phi}(\mathbf{x}) - \tilde{\boldsymbol{\Phi}}(\mathbf{x})\|^2 = \|\boldsymbol{\Phi}(\mathbf{x})\|^2 + \|\tilde{\boldsymbol{\Phi}}(\mathbf{x})\|^2 - 2\boldsymbol{\Phi}^\top(\mathbf{x})\tilde{\boldsymbol{\Phi}}$$

where

$$\|\boldsymbol{\Phi}(\mathbf{x})\|^2 = \mathbf{k}(\mathbf{x}, \mathbf{x})$$

$$\|\tilde{\boldsymbol{\Phi}}(\mathbf{x})\|^2 = \mathbf{N}\lambda_\mathbf{k}(\mathbf{k}^\top(\mathbf{x})\boldsymbol{\alpha}_\mathbf{k})^2$$

$$\boldsymbol{\Phi}^\top(\mathbf{x})\tilde{\boldsymbol{\Phi}}(\mathbf{x}) = (\mathbf{k}^\top(\mathbf{x})\boldsymbol{\alpha}_\mathbf{k})^2$$

hence

$$\|\boldsymbol{\Phi}(\mathbf{x}) - \tilde{\boldsymbol{\Phi}}(\mathbf{x})\|^2 = \mathbf{k}(\mathbf{x}, \mathbf{x}) + (1 - 2\mathbf{N}\lambda_\mathbf{k})(\mathbf{k}^\top(\mathbf{x})\boldsymbol{\alpha}_\mathbf{k})^2$$